Genome-scale Bacterial Transcriptional Regulatory Networks: Reconstruction and Integrated Analysis with Metabolic Models

José P. Faria, Ross Overbeek, Fangfang Xia, Miguel Rocha, Isabel Rocha, Christopher S. Henry

José P. Faria
University of Minho, Institute for Biotechnology and Bioengineering/Centre of Biological Engineering, Braga, Portugal
Argonne National Laboratory, Mathematics and Computer Science Division
Argonne, Illinois, United States
jplfaria@deb.uminho.pt

Ross Overbeek
Fellowship for Interpretation of Genomes
Burr Ridge, Illinois, United States
ross@TheFIG.info

Fangfang Xia
Argonne National Laboratory, Mathematics and Computer Science Division
Argonne, Illinois, United States
fangfang@cs.uchicago.edu

Miguel Rocha
University of Minho, Department of Informatics/CCTC, Braga, Portugal
mrocha@di.uminho.pt

Isabel Rocha
University of Minho, Institute for Biotechnology and Bioengineering/Centre of Biological Engineering, Braga, Portugal
irocha@deb.uminho.pt

Christopher S. Henry
Corresponding Author
Argonne National Laboratory, Mathematics and Computer Science Division
Argonne, Illinois, United States
cshenry@mcs.anl.gov

Abstract

Advances in sequencing technology are resulting in the rapid emergence of large numbers of complete genome sequences. High-throughput annotation and metabolic modeling of these genomes is now a reality. The high-throughput reconstruction and analysis of genome-scale transcriptional regulatory networks represent the next frontier in microbial bioinformatics. The fruition of this next frontier will depend on the integration of numerous data sources relating to mechanisms, components, and behavior of the transcriptional regulatory machinery, as well as the integration of the regulatory machinery into genome-scale cellular models. Here we review existing repositories for different types of transcriptional regulatory data, including expression data, transcription factor data, and binding site locations; and we explore how these data are being used for the reconstruction of new regulatory networks. From template network-based methods to *de novo* reverse engineering from expression data, we discuss how regulatory networks can be reconstructed and integrated with metabolic models to improve model predictions and performance. We also explore the impact these integrated models can have in simulating phenotypes, optimizing the production of compounds of interest, or paving the way to a whole-cell model.

Keywords

Genome-scale metabolic model (GSM); Transcriptional Regulatory Network (TRN); *De novo* reverse engineering; Integrated Metabolic and Regulatory Models

Introduction

Systems biology has provided numerous tools for modeling biological systems [1], many of which depend on the reconstruction of genome-scale metabolic models (GSM). These models now exist for a growing number of organisms, including prokaryotic, archaeal, and eukaryotic species [2]. With the advent of next-generation sequencing, the development of GSMs has become routine [2, 3], and many steps involved in the reconstruction and optimization of draft GSMs have been automated [4]. Algorithms and methods for GSM reconstruction have been reviewed in detail elsewhere [5-7].

However, nearly all existing GSMs fail to account for the impact of gene expression regulation on metabolic activity. In order to capture the impact of regulation on the behavior of an organism, a GSM must integrate some abstraction of regulatory mechanisms, which include the activity of RNA polymerase, transcription factors (TFs), promoters, transcription factor binding sites (TFBS), and sigma factors. Sigma factors allow the recognition of the enzyme by the promoter region, enabling transcription to begin. TFs bind to specific TFBSs in the promoter region and can act as activators, repressors, or both (dual regulators). In eukaryotes, TFs are able to perform other tasks affecting regulation, such as chromatin-modifying activities [8]. Other elements have been identified as taking part in the control of transcription regulation in bacteria, such as riboswitches [9], RNA swiches [10], antisense RNA [11], or microRNAs [12]. Here we focus on regulation by transcription factors, a mechanism illustrated in Fig. 1. Also displayed are some of the technologies, tools, and resources necessary for reconstructing transcriptional regulatory networks.

*** Insert figure 1 around here ***

The integration of these regulatory mechanisms in GSMs requires methods for the reconstruction and analysis of transcriptional regulatory networks (TRNs). Once a regulatory model has been constructed for an organism, it can be integrated with GSMs to improve predictive accuracy and reveal new biological insights. For example, some cellular processes exhibit a dominance of regulatory mechanisms, affecting their behavior and leading to incorrect predictions when only metabolism is accounted for [13]. The first genome-scale integrated metabolic and regulatory model for *E. coli* [14] revealed that regulation significantly affects growth phenotype predictions, and these predictions improved with the addition of regulatory constraints. Simultaneously, the study of TRNs has unveiled novel interactions; in *Salmonella enterica,* 14 regulators were identified that affect the same genes leading to a systemic infection [15]. Similar studies led to the discovery of novel regulatory mechanisms in *Saccharomyces cerevisiae* [16].

Here, we review the reconstruction of TRNs and their integration with metabolic models. First, we explore the data available for TRN reconstruction, covering the most prominent databases of expression data and repositories of TF/TFBS data. Next, we examine how data availability triggered the development of a variety of TRN inference methods, including reverse engineering from expression datasets [17-21], network inference from TFBS site data [22-24], and knowledge-based template methods [25].

The integration of regulatory and metabolic networks for predictive modeling is possible only with the development of integrated phenotype simulation methods. The most widely used approach for simulating GSMs is flux balance analysis (FBA) [26]. To account for regulatory information, FBA was expanded with new methodologies, including rFBA [13] and SR-FBA [27]. We review these FBA-based methodologies, as well as other approaches that allow for a characterization of alternative cellular states [28] and for the integration of omics data [29, 30].

Regulation data for TRN reconstruction – From standards and technologies to databases

The development of microarray technologies gave rise to a revolution in biomedical research [31], also bringing new problems such as quality control of experiments [32] and selection of an appropriate level of detail [33]. To address these issues, the Functional Genomics Data Society (FGED) launched a proposal to standardize the publishing and sharing of microarray data (MIAME) [34]. The majority of the community adopted the proposal, requiring authors to follow the MIAME guidelines. Publishers also required authors to store data [35] in either NCBI's Gene Expression Omnibus (GEO) [36] or EBI's ArrayExpress [37], the major public gene expression data repositories, both MIAME compliant.

These databases integrate data from a variety of technologies that can help determine regulatory interactions, although expression profiling and genome binding and occupancy studies have become the most prevalent. Expression profiling techniques vary from the traditional array oligonucleotide hybridization technology for measuring gene expression level to mRNA quantification methodologies, such as serial analysis of gene expression (SAGE) [38, 39] or reverse transcriptase PCR (RT-PCR). Genome binding and occupancy experiments have the advantage of identifying the spots corresponding to DNA-protein binding targets. Chromatin immunoprecipitation with array hybridization (ChIP-chip) [40, 41] is used to overcome limitations of common expression profiling. Other ChIP technologies have also been developed in combination with different expression techniques such as SAGE (ChIP-SAGE [42]) to achieve a particular level of detail, depending on the organism and tissue studied [43]. With the development of next-generation sequencing technologies, ChIP-Seq [44] and RNA-Seq emerged [45, 46]. ChIP-Seq enables whole-genome ChIP assays, while RNA-Seq provides a capacity for direct measurement of mRNA, small RNA, and noncoding RNA abundances [47]. ChIP methods have been widely used to collect expression data from *E. coli* [48-50]; and, more recently, RNA-Seq methods have been adjusted for studying bacterial transcriptomes [51, 52]. RNA-Seq has been also successfully used to detect transcription start sites [53] that can be used for regulon inference.

Data available for TRN inference can be categorized into two major groups: (i) databases of gene expression data (including genome binding experimental data), and (ii) databases of TF and TFBS. Table 1 shows the most notable databases of the former group.

*** Insert Table 1 around here ***

We surveyed GEO, as the major expression database, gathering statistics on the type of studies conducted, availability of data, quantification of bacterial data, and the most represented microbes (Fig. 2). These statistics clearly indicate that most of the current data are from expression profiling, with 18,498 experimental series (85%). Although next-generation sequencing technologies were introduced recently [54], we can already see a change in the types of experiments being performed (Fig. 2b). Examining the organisms for which expression data are available, we find that only 7% of datasets are from bacteria (Fig. 2c), with *Escherichia coli* being the most represented prokaryote (Fig. 2d).

*** Insert Figure 2 around here ****

Table 1 also includes other notable databases, from which we highlight the Many Microbe Microarrays Database (M3D) [55] currently holding around 2,000 microarrays for *Escherichia coli*, *Saccharomyces cerevisiae,* and *Shewanella oneidensis*. The data available are all from Affymetrix single channel microarrays, allowing a uniform normalization procedure and higher-quality data. The *E. coli* data have already been applied for TRN inference [56].

*** Insert Figure 3 around here ***

Figure 3 shows the discrepancy between the number of sequenced genomes and the number of genomes for which any type of expression data exists. In this study, we cluster bacterial genomes available in the PubSEED [57] (a large repository of genomes and annotations) at the taxonomical level of family. The set of 20 bacterial families associated with expression data in GEO are shown in the phylogenetic tree. On average, 16.2% of the 3,493 PubSEED genomes that fall into these families have expression data linked to them. Expression data are available for 55% of the genomes in the Gammaproteobacteria family, demonstrating the extensive amount of data available for this taxonomic clade. In contrast, more than half of the bacterial phyla have expression data for less than 10% of their species, revealing that numerous phylogenetically distinct clusters of microbes have little gene expression experimentally characterized.

*** Insert Table 2 around here ***

Repositories with regulatory interactions also hold valuable information. Table 2 shows the most comprehensive resources available for prokaryotes. Organism-specific databases are available for well-known organisms such as *E. coli*, *B. subtillis,* and *M. tuberculosis*, including a comprehensive collection of regulatory information. Among those, RegulonDB is the most comprehensive resource for regulatory interactions data of any single organism (*E. coli*). In its latest release, genetic sensory response units are introduced to better represent the biology of gene regulation [58], trying to capture all the phenomena involved in regulation, from the initial signal to gene response. Another major resource for *E. coli* data is EcoCyc [59], integrating RegulonDB and curated data from over 21,000 publications and TRN descriptions that include genes, ligands, and regulators with their targets. DBTBS [60] is the major resource for *B. subtillis* regulatory data.

Less comprehensive databases present fewer types of different regulatory information (sometimes only TFBS predictions or TF information) but cover a wide range of bacteria (Table 2). Notable examples are ODB [61], which stores known operon data for about 10,000 operons in 56 organisms and putative operons for over 1000 genomes; RegTransBase [62], which collects regulatory data from the literature; and RegPrecise [63], a repository of manually curated regulons that provides tools for regulon propagation.

Reconstruction of TRNs can use different types of data, and the accurate selection of data/database(s) for the method of choice is paramount in the reconstruction process. Organism-specific databases are particularly useful for reverse engineering methodologies as training datasets and essential for validation. Methodologies based on comparative genomics approaches make good use of less comprehensive databases but cover a wider range of organisms.

TRN Reconstruction – From template networks and inference algorithms to integration with GSMs

TRN reconstruction aims to make sense of gene expression and binding site data by revealing the interactions between the different elements of the cell's regulatory machinery. Different methodologies have been proposed for TRN inference. However, there is no consensus for classification in the literature. Some reviews classify methods as bottom-up and top-down [64], others focus on inference from a specific type of data such as gene expression [65], while others present methods and computational tools [66]. Here, we review and categorize different methodologies within two major types: genomics-driven and data-driven. The first uses comparative genomics approaches, while the second refers to *de novo* reverse engineering from expression data. Within the genomics-driven approaches, we present two methodologies: template network-based methods and TFBS data-based methods via prediction of *cis*–regulatory elements, including propagation from known regulons and *ab initio* regulon inference. The comparative genomics approaches are described in Fig. 4a and 4b; Fig. 4c describes data-driven methods from expression data.

*** Insert Figure 4 around here ****

Template network-based methods

Template-based methods [67] rely on one or more well-characterized networks to serve as a starting point for the reconstruction. These methods exploit the conservation of prokaryotic gene networks [68-71] to reconstruct TRNs (Fig. 4a). Starting with a well-characterized network, a search for orthologous genes (e.g. using bidirectional best hits [72]) is conducted on the genome of interest. With the orthologous TFs and their targets noted on the target genome, random networks are generated from the template network to confer statistical strength to the new reconstructed interactions in the target genome, since this shows the significant trends. After this analysis, the new interactions on the target genome are reconstructed. This approach can be useful for propagation of TRNs to other strains of a model organism or to closely related organisms.

This methodology presents some limitations, however. The first is intrinsic: the need for a high-quality template network derived for an organism that is phylogenetically close to the organism being studied. A long phylogenetic distance between the template and the target organisms can generate meaningless interactions; hence the choice of the template network is of paramount importance for the reconstruction. Another limitation is the scale of the network to be reconstructed; here our focus is genome-scale network reconstruction, and reconstructions on this scale depend on the availability of a template network that also exists at the genome scale.

TFBSs data-based methods via prediction of *cis* – regulatory elements

TRN reconstruction from binding site data can also be defined as a comparative genomics approach. Prior to the development of the first binding-site approaches, most methods relied almost entirely on functional information from expression data [19, 73]. The GRAM (Genetic Regulatory models) algorithm [74] was the first to combine the use of expression data and binding site data in a genomewide inference process, enabling the inclusion of information about physical interactions between regulatory genes and their targets. Other work focused on the conservation of the regulatory machinery across different organisms.

Regulogger [75] was introduced to generate *regulogs*, or sets of genes that are co-regulated and have their regulation processes conserved across several organisms. Using *Staphylococcus aureus*, regulogs were produced for well-known sets of genes and provide clues about the functions of unannotated genes. Studies of δ-proteobacteria [23] revealed that very diverse species of proteobacteria have similar regulatory mechanisms.

The principles behind this methodology were reviewed by Rodionov [76]. Figure 4b describes one of the two strategies proposed. The first step is to gather all available information related to TFs and TFBSs in a selected model organism. These data are then used as a training set for the TFBS model. The accuracy of the methodology is closely connected to the quality and quantity of sequences used for training. *E. coli* is usually used as a model species for gram-negative bacteria, and *B. subtilis* for gram-positive bacteria. If the TFBSs corresponding to a particular TF are unknown, all genes regulated by the TF in the model species are identified, and then orthologs for these genes in closely related genomes are found. With a TFBS training set built by this process or experimentally determined (see Table 2), positional weight matrices (PWMs) are constructed for the collection of binding sites. Several algorithms are available that perform motif pattern recognition [77] to construct PWMs. One of the first algorithms developed for this task was AlignACE [78]. This algorithm was recently upgraded to W-AlignACE [79] incorporating a new learning approach [80] and showing increased accuracy in obtaining PWMs for gene sequences, gene expression data, and ChIP-chip data [79]. Using the PWMs, one can perform a genomewide search for putative TFBSs on the target genomes.

This comparative-genomics-based approach requires a high-quality training set; using genomes that are not closely related can lead to generation of false positive TFBS predictions. Even for a set of closely related genomes, selecting a threshold for binding site detection can be difficult. The final step of the TFBS prediction involves the verification of site consistency. Early studies on *E. coli* and *H. Influenzae* regulon predictions showed conservation of co-regulated genes by orthologous TFs [81]. Based on this principle, a search is conducted for binding sites upstream from the operons regulated by each TF. If the site is conserved, the TFBS prediction is assumed to be correct. On the other hand, if matches to the predicted TFBS motif are found dispersed across the genome, the prediction is assumed to be a false positive. By accounting for changes in the operon structure, further consistency checks are possible. This method showed improved results in binding site detection in several studies such as nitrate and nitrite respiration in γ-Proteobacteria [82] and nitrogen metabolism in gram-positive bacteria [83].

These methodologies have been implemented in the RegPredict web resource [84], a state-of-the-art tool for TRN reconstruction with TFBS data. The webserver comprises a large set of comparative genomics tools available in two reconstruction frameworks; the first reconstructs regulons for known PWMs, and the second performs *de novo* regulon inference for unknown binding sites using analysis of regulon orthologues across closely related genomes. One of the novelties of RegPredict is the concept of CRONs (Clusters of co-Regulated Orthologous Operons) to facilitate and improve consistency check. This semi-automated approach provides the community with a more swift reconstruction, curation and storage of regulons. RegPredict was used for TRN reconstruction of the central metabolism of the *Shewanella genus* [85], for the analysis of the regulation of the hexunorate metabolism in Gammaproteobatceria [86], and for the elucidation of control mechanisms for proteobacterial central carbon metabolism by the HexR regulator [87]. FITBAR [88] is another web tool for prokaryotic regulon prediction that aims to fill the gap of the lack of statistical comparison for calculating the significance of the predictions.

Techniques also exist for predicting TFBSs when the available regulatory information is not sufficient for regulon-based approaches. Phylogenetic footprinting [89] identifies highly conserved untranslated regions (UTRs) upstream from the genes of interest, since these are prime regulatory site candidates. An orthologous search for these regions is performed across closely related genomes; candidate binding sites are identified; and these sites are used to perform a regulatory motif search across all analyzed genomes. This technique successfully identified the FabR regulon in *E. coli* and regulon members in several cyanobacteria genomes [90]. Another approach has been described as subsystem oriented [76] based on the hypothesis that one TF regulates the genes on the same metabolic pathway. A search for orthologous genes on the same metabolic pathway of closely related genomes is conducted. Using the orthologous operons from the same subsystem, one can perform a motif search to build the PWM and search for TFBS. Concepts of this approach were also implemented in RegPredict with the introduction of the SEED subsystems [57] for regulon reconstruction and curation.

De novo reverse engineering

As gene expression data became available through microarray technologies, development began on methods for inference of regulatory networks from expression data [91]. Early reviews describe several mathematical formalisms such as Bayesian networks, Boolean networks, and differential equations to represent regulatory networks [92], together with appropriate algorithms to support network inference.

The development of these methodologies led to the creation of the DREAM (Dialogue for Reverse Engineering Assessments and Methods) project in 2007 [93], bringing together experts from different areas and aiming to provide tools to enable the unbiased evaluation of various methods [94], hosting annual challenges. The lessons gained from the results obtained in those challenges have provided improved methods for network inference [95]. Each year different methods are ranked as top performers on specific sub challenges that differ in either the type of data or network size.

Past reviews have categorized reverse engineering network inference methods according to (i) mathematical modeling approach [65, 96], (ii) module-based or direct inference methods [64, 97], and (iii) unsupervised and (semi)-supervised methodologies [64, 98, 99].

In the first category [65, 100], the differential equation (ODEs)-based [101, 102], mutual information-based [103, 104], and Bayesian network-based methods [105, 106] are the most popular approaches. Other notable approaches are based on Boolean networks [107], neural networks [108, 109], correlation analysis [110], and relevance networks [111].

The second category divides methods into those based on a modular view of regulatory networks that infer regulatory programs for sets of co-expressed genes and those able to infer the regulatory behavior of individual genes (direct inference) [79]. Module-based inference is inspired by evidence that regulatory networks exhibit a modular structure of co-expressed genes [112, 113], using a separate algorithm for the module inference step, typically based on clustering or biclustering algorithms, such as cMonkey [114]. Direct inference methods search for single interactions between targets and their regulators [56, 115] (Fig. 5). A comparison between representative methods of both approaches showed that none can be defined as the best solution [97]: the module-based method LeMoNe [116, 117] is able to retrieve more efficiently targets for regulators with a high number of targets, and the direct-inference method CLR [56] is preferable for detecting regulators with one or few targets. Thus, these methods can be seen as complementary when handling genome-scale regulatory model reconstruction.

*** Insert Figure 5 around here ***

The third category divides methods into supervised [118, 119] and unsupervised [120, 121]. The former use a training set of known interactions creating classification problems (e.g., to infer whether a given gene is regulated by a transcription factor) (Fig. 5). Some supervised methods are known as semi-supervised [122, 123]. Supervised methods have shown to provide more accurate predictions than unsupervised methods [124], with successes in expanding the compendium of TF-gene interactions in *E. coli* [122]. At the same time, when inferring interactions for an organism that is not well known, the lack of a proper training set can lead to a better performance by unsupervised methods.

A detailed review of the mathematical formalisms and detailed inference algorithms is out of the scope of this review. From the overwhelming number of methods available, we chose to briefly describe 10 methods, including the most widely used, the most recent [64], and the best performing from the DREAM challenges [94, 95, 125-127]. We focus our review on methods that produce genome-scale regulatory network reconstructions in the form of regulatory models that may be integrated with GSMs. While no method currently exists that completely satisfies these criteria, several algorithms, given in Table 3, can provide important results in the route to achieve the goal of fully integrated genome-scale models.

*** Insert Table 3 around here ***

ARACNE [115] is one of the most widely used methods, first applied to infer regulatory interactions on human B cells [128]. Also, it has shown capacity for genome-wide inference in bacterial species such as *Streptomyces coelicor* [129]. CLR (context likelihood of relatedness) introduced the use of data from different experimental conditions for the same organism to infer regulatory interactions and enabled the identification of over 700 novel interactions in *E. coli* [56]. Being one of the most cited methods with an ability to predict edges in the RegulonDB, CLR is the method of choice for regulatory interactions studies [130]. It was recently used to unveil virulence factors in *Salmonella* [131]. A newer algorithm based on CLR, called SA-CRL (synergy augmented-CLR) [132], was the best-performing method in the

DREAM2 genome-scale inference challenge, exploiting the concept of synergy among multiple interacting genes [133], where a pair of genes is used to infer the expression of a third to increase prediction accuracy.

The Inferelator [134] was applied for genomewide reconstruction of *Halobacterium*. A mixed approach combining this method with CLR was one of the top performers in the DREAM3 *in silico* network challenge [135], using a modified version of CLR to compute mutual information values that are subsequently used by Inferelator to produce an ODE model. This method, called tlCLR (time-lagged CLR), takes advantage of two types of data: steady-state data from knockout experiments and time series gene expression data. Another method using different types of data was introduced by Yip *et al*. [136] gathering steady-state data from a noise model and time series data from an ODE model; this method was the top performer of the DREAM3 *in silico* challenge. Most algorithms in Table 3 can use steady-state or time series data, thus showing the benefits of integrating both types of data.

DREAM5 featured a genome-scale network inference challenge with a large dataset from a compendium of microarray data for *E. coli* comprising 805 chips, 334 TFs, and 4,511 genes. Large datasets were also provided for network inference on *Saccharomyces cerevisiae* and *Staphylococcus aureus*. GENIE3 (GEne Network Inference with Ensemble of trees) [119] uses tree-based methods [137] decomposing the inference problem of *p* size into *p* distinct regression models. This method was the best performer overall and the top performer in the *in silico* network. GENIE3 had already been the best performer in the DREAM4 *in silico* inference for the 100-gene-multifactorial subchallenge, where only multifactorial data were provided, and showed equal capacity in successfully inferring networks from real data when compared was widely used methods such as CLR and ARACNE [119].

Several methods integrate multiple data types (e.g., inference from expression, binding site data) to facilitate TRN reconstruction. SEREND (SEmi- supervised REgulatory Network Discoverer) [138] uses a semi-supervised and iterative approach to unveil regulatory interactions. SEREND depends on a curated set of TF-gene interactions and TF-gene motif scores as a training set to construct a logistic regression model. The known predictions are then expanded and the predictions validated with ChiP-chip and time-series expression data. This approach was used to better predict and to give new insights into the factors involved in activation and repression in the aerobic/anaerobic regulation mechanism in *E. coli* [138].

GPS (Gene promoter Scan) [139] is also able to integrate other types of data; but as a module-based method, it follows a different approach. GPS is a machine learning method that builds promoter models and their relationships computed from a dataset. In the next step, characterized profiles (groups of promoters) are generated. The best profiles are used as candidates for genomewide predictions. Studies with *E. coli* and *S. enterica* using GPS unveiled previously unknown interactions and novel members of the PhoP protein controlled regulon [139].

DISTILLER [140] is another method that exploits the concept of regulation modularity integrating other sources of data for network inference. This framework can be applied to any organism and incorporate motif and ChiP-Chip data. The integrated approach was used to study the *FNR* regulon in *E. coli* identifying novel predictions that were experimentally validated. These studies provided insights on modularity dynamics pointing to the existence of polycistronic transcription [141].

A search for the best inference method usually turns to benchmarking studies; but the choice of benchmark datasets presents a problem, with different studies showing very sparse results [142, 143]. Lessons from all the DREAM challenges show that there is no individual best method. Results from community predictions, a combination of several reverse engineering methods, are closer to a state-of-art/best method, outperforming results from individual algorithms. The determination of error profiles enables the advantages and limitations of each inference method to be assessed in order to determine which method is "the best" for a specific inference problem.

The methods described above show recent advances, providing a good summary of the huge number of approaches that have been put forward. However, the underlying problem is complex, given the large search spaces involved and the still restricted availability of data that leads to an undetermined problem where many solutions can explain the data equally well. Hence, most of the methods rely on heuristic methods using different strategies to simplify the problem. The most important simplification is to reduce the search for a network or model explaining the data, with a huge number of possible interactions between the different entities involved, to the search of individual interactions or to small clusters or modules. This allows in some cases for distinct methods to be integrated to better support the results and, in the most elaborate methods, being followed by steps of determining regulatory programs based on these individual interactions.

Phenotype simulation by integrated metabolic and regulatory networks

The simulation of phenotype from genotype using reconstructed models has been one of the major goals and challenges of systems biology [144-146]. Early work on the integration of metabolic networks with gene expression data revealed that some cellular phenotypes cannot be described by the metabolic flux distribution alone [147]. Whole-cell modeling is required to capture many phenotypes, and while this has been one of the great challenges of the century [148], integration of regulatory networks is one key milestone toward achieving this goal [149]. Significant advances have been made in the reconstruction of metabolic, regulatory, and signaling networks [150, 151], as well as in the integrated simulation of these three network types [152, 153]. Here, we focus on the potential for the simulation of integrated metabolic and regulatory networks and the challenges that arise in this integrated approach [154].

*** Insert Figure 6 around here ***

Several mathematical formalisms have been applied to model different types of biochemical networks (e.g., Boolean and Bayesian networks, constraints-based optimizations, ODEs). The many types of approaches for integrated network reconstruction and analysis have been reviewed recently [66, 155, 156]. Here, we focus on the methods that can be applied at genome scale, mainly stoichiometric models using the constraints-based approach [157, 158].

Constraints-based stoichiometric models do not account for intercellular dynamics. Instead, they assume a pseudo-steady-state for the cell, in which metabolite accumulation does not occur. This is described mathematically by a set of linear constraints on the flux through each metabolic reaction, defined by the mass balance for each internal metabolite (Fig. 6):

$$\boldsymbol{S} \,.\, \boldsymbol{v} = \boldsymbol{0},$$

where $\boldsymbol{S}$ represents the stoichiometric matrix and $\boldsymbol{v}$ the vector of fluxes through all metabolic reactions. The set of fluxes that satisfy these constraints define the feasible space for reaction fluxes (Fig. 5). Constraints can be imposed on reaction reversibility and directionality (***v* > 0),** on enzyme capacity (**v < v$_{\mathbf{max}}$**), and on nutrient availability and uptake.

Extensions have been made to these basic mass-balance and flux boundary constraints to capture the additional constraints imposed by regulatory interactions. Figure 7 shows existing methods for analysis and simulation of integrated metabolic and regulatory networks. Global network analysis methods such as extreme pathway analysis [159] were developed to analyze specific pathway properties, such as length and redundancy. These methods were used successfully to characterize changes in the solution space with the addition of regulatory constraints [28].

*** Insert Figure 7 around here ***

The flux balance analysis approach uses linear programming to identify the specific flux distributions that satisfy problem constraints and best reflect the state of the cell or represent target states for metabolic engineering [160, 161]. FBA was expanded to account for regulatory information with the introduction of rFBA (regulatory FBA) [13], which uses a Boolean logic formalism to define additional constraints specifying which genes in the network are ON or OFF, based on specified stimuli (e.g., stress). This approach was successfully applied with the first genome-scale integrated model of metabolism and regulation in *E. coli*, resulting in the correction of several phenotype predictions compared with the use of mass balance and flux boundary constraints alone [14]. However, this approach requires the integrated model to be initialized at a relevant state for the regulatory components of the system. The Boolean regulatory constraints are then applied to determine how the state of the regulatory components will change over time in response to stimuli. Selection of a relevant initial condition for the model remains a challenge for this methodology, since many equally consistent states exist for a set of stimuli, with equally valid associated flux distributions.

To address some of the limitations of rFBA, SR-FBA (Steady-State Regulatory FBA) [162] was introduced, differing from rFBA in that it accounts for metabolic and regulatory constraints in a single step and quantifies the impact of these constraints on the flux distribution. This methodology enables the rapid exploration of feasible combined regulatory and metabolic states, and it rapidly identifies constraints that are internally inconsistent, preventing their simultaneous enforcement in a single steady-state. Yet, therein lies the substantial limitation of this approach, since inconsistent regulatory constraints often arise, because regulatory mechanisms exist to manage transitions between states of the cell in response to stimuli. Some of these transitions involve a cascade of intermediate unstable states that cannot be captured by the SR-FBA formalism. The constraints that manage these cascade transitions are not designed to be simultaneously enforced with all other constraints in the cell, meaning they appear to be internally inconsistent.

The quest for a whole-cell model led to the development of methods that also integrate signaling networks. Two methods have been proposed: iFBA (integrated FBA [152]) and idFBA (integrated dynamic FBA) [153]. iFBA is an expansion of the rFBA approach that aims to integrate signaling models, when available, for an organism or pathway being studied. An rFBA model for the central metabolism of *E. coli* [163] was combined with an ODE kinetic model for the phosphotransferase system, showing improved predictions compared with both rFBA and ODE models. The novelty of idFBA is the incorporation of slow and fast reactions in the stoichiometric framework. Slow reactions are incorporated directly into the stoichiometric matrix with a time delay; fast reactions rely on the pseudo steady-state assumption of the FBA approach. idFBA was applied to the analysis of yeast metabolism [164], demonstrating an approximation for the time-course prediction of time-delayed reactions, with the advantage of requiring fewer measured parameters than with full kinetic modeling.

Before methods such as rFBA, srFBA, iFBA, or idFBA can be applied, TRNs must be translated into Boolean network models that connect external stimuli to internal metabolic reaction activity. The PROM (Probabilistic Regulation of Metabolism) [165] approach was introduced to avoid the translation to Boolean constraints by enabling the generation of integrated models directly from high-throughput TRN data. PROM aims to circumvent the Boolean approaches that would consider a gene as either ON or OFF, with results outperforming rFBA. The differences in the predictions are attributed to the Boolean formalism of rFBA, which establishes a set of "rigid" flux restrictions, where PROM presents a more continuous flux restriction. The reconstruction of an integrated model for *M. tuberculosis* showed a potential use of PROM for drug target prediction. PROM can be seen as the closest methodology for semi-automated reconstruction of integrated metabolic and regulatory networks.

Transcriptional controlled FBA (tFBA) [166] is another method that uses experimental expression data for the assessment of the regulatory state. Like PROM, tFBA aims to surpass the rigid ON/OFF gene states of a purely Boolean formulation by introducing the concept of more relaxed up/down constraints. As more experimental data are available, the level of expression of a gene can be observed to change under specific conditions without being entirely shut off. This method shows how the addition of large quantities of expression data can provide a way to improve FBA-based methods in the absence of kinetic parameters for metabolites and reactions.

Discussion

In this survey, we begin with an overview of the data currently available for TRN reconstruction, revealing the limited number of datasets available for bacterial organisms, despite the massive amount of existing microarray data (Figs. 1 and 2). We demonstrate through a phylogenetic analysis of the available expression data that large numbers of diverse organisms for which reference genomes are now available have never been examined using transcriptomic techniques. In order to fully understand bacterial regulation, expression data must be collected under a variety of conditions for as many diverse genomes as possible. We also show how next-generation sequencing technologies are beginning to dominate the latest submissions to the gene expression data repositories. While these new technologies enable the community to collect more data at a faster and cheaper price, they face the familiar problem of data standardization. Recent studies show how widespread batch effects, such as laboratory conditions, technicians, and reagent brands lead to incorrect analysis of data and different results across different laboratories [167].

As for data relating to the regulons, transcription factors, binding sites, and stimuli that comprise the TRN itself, comprehensive databases are available for a few specific organisms. Multiorganism databases do exist, but these typically focus on one type of regulatory information, lacking the information needed to fully capture the regulatory effects. The latest version of RegulonDB makes an effort in the direction of representing the complete regulation by introducing genetic sensory response units.

Next, we examine how the methods applied for the reconstruction of TRN have progressed over the past decade. As the number of available reference genomes with expression data has increased, we see a corresponding increase in the number and power of approaches based on comparative genomics. With the increasing amounts of consistent high-quality expression data, we are also seeing increasing success with methods based on the reverse engineering of TRN from expression data. As these two examples amply demonstrate, the best method for TRN reconstruction depends on the amount and type of data available. Although the size of the desired TRN to be inferred is also an important factor, we suggest that genome-scale networks will always be desired in the near future. We also note the success of community efforts that combine the advantages of several reconstruction approaches, showing that hybrid approaches are the most successful given the present knowledge, where the complementary nature of the approaches helps to improve accuracy.

In the final portion of our review, we examine several approaches for the reconstruction and analysis of integrated metabolic and regulatory models. These approaches have been successfully applied to improve our ability to accurately predict phenotype from genotype, to explore the impact of regulation on the metabolic pathways, and to simulate regulatory interactions that are continuous rather than discrete. Industrial successes in fields such as bioethanol production show the potential of current models and importance of improving these models [168, 169]. Adding a “layer” of regulation can help unveil and predict unobserved phenotypes. Strain optimization has been one of the main objectives of metabolic engineering, and the potential for improvements integrating regulatory information recently led to the development of methods that account for this type of information [170, 171]. Yet, we still lack a full understanding of the interplay between regulation and metabolism. Several studies have shown how major transcriptional changes are not always followed by changes in the metabolic flux [172, 173].

Several unknowns remain in the analysis and reconstruction of integrated biochemical networks, mostly because we do not yet possess a full understanding of regulation. For example, some efforts have been made to develop methods to account for metabolic activity effects regulated by post-transcriptional effects [174]. Methods such as PROM and tFBA allow the relaxation of constraints to try to account for regulatory effects. Even with transcriptional regulation, there are biological effects that these network models fail to reproduce. For example, chromosome structure can physically constrain bacterial transcriptional regulation [175]. Epigenetics of transcriptional regulation are also difficult to account for, and some chemical marks have been described to be linked to this type of mechanism in bacteria [176].

Some of the methods described rely on basic assumptions such as that the same TF regulates orthologous genes or that the same TF may regulate genes in the same pathway. These assumptions may fail to represent reality, however, since TRNs show considerable plasticity in their structure. Orthologous regulators have been shown to control different pathways across different species [177], and global regulators have been shown to regulate different mechanisms [178]. Incorporation of models in evolutionary processes such as duplication and horizontal gene transfer has been proposed to deal with TRN plasticity [76, 179]. TRN also showed stochasticity [180], which can be an issue, especially when these networks are modeled by using a Boolean formalism that further propagates these stochastic effects [181].

Most recently Karr *et al*. [182] introduced the first whole-cell model for *Mycoplasma genitalium*. Integrating 28 submodels, the authors managed to validate the model across a wide set of experimental data, pointing out its potential for novel biological discovery in *M. genitalium*. It must be noted however, that *M. genitalium* is the smallest bacterial genome, with only 525 genes. Thus, while this methodology does represent a large step forward toward the goal of true whole-cell models, much more work must be done before similar models can be constructed for larger and more complex organisms.

As the pursuit of a whole-cell model continues, we expect novel regulatory interactions will be discovered in our drive to build a full understating of cell regulatory machinery.

Key Points

1.) Large numbers of phylogeny for which genome sequences are available still lack any gene expression data.
2.) Repositories of data on TRN tend to be comprehensive organism specific or narrowly focused multiorganism.
3.) The best methods for reconstruction of TRN from data depend on the size of the desired network and the types/amount of data available; but, in general, hybrid methods that combine many approaches produce the best results.
4.) Methods for integrating regulatory and metabolic models must include both steady state and dynamic components, and they must accommodate more than just Boolean regulation in order to fully capture the behavior of transcriptional regulation.
5.) Integration of regulatory constraints in genome-scale metabolic models results in substantial improvements in accuracy of phenotype predictions, particularly since many phenotypes cannot be fully explained without accounting for regulation. Yet, some regulatory mechanisms still exist that are poorly understood and require further study.

Acknowledgments

Faria acknowledges funding from grant FRH/BD/70824/2010 of the FCT (Portuguese Foundation for Science and Technology) PhD program. The work was supported in part by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), National Funds through the FCT within projects FCOMP-01-0124-FEDER-015079 (ToMEGIM - Computational Tools for Metabolic Engineering using Genome-scale Integrated Models) and FCOMP-01-0124-FEDER-009707 (HeliSysBio – molecular Systems Biology in Helicobacter pylori), the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357, and the National Science Foundation under grant number 0850546.

Authors biographical note

José P. Faria is a Ph.D. student in the Centre of Biological Engineering at the University of Minho, Portugal. He conducts part of his research at Argonne National Laboratory.

Ross Overbeek is a founding fellow of the Fellowship for Interpretation for genomes (FIG) and a scientist at Argonne National Laboratory. He is an expert in bioinformatics and genome annotation.

Fangfang Xia is an assistant computer scientist at Argonne National Laboratory. He is an expert in bioinformatics, phylogenetics, and high-performance computing.

Miguel Rocha is an assistant professor in the School of Engineering, University of Minho, Portugal, where he also leads the CCTC research center. His main research interests lie in the fields of bioinformatics, machine learning, and evolutionary computation.

Isabel Rocha is an assistant professor at the Institute for Biotechnology and Bioengineering at the University of Minho, Portugal. She conducts her research in the fields of metabolic engineering and systems biology.

Christopher S. Henry is an assistant computational scientist with appointments at Argonne National Laboratory, the University of Chicago, and Northwestern University. He is an expert in the reconstruction, optimization, and analysis of genome-scale metabolic models.

Bibliography

1. Chuang HY, Hofree M, Ideker T. A decade of systems biology, Annu Rev Cell Dev Biol 2010;26:721-744.
2. Reed JL, Famili I, Thiele I et al. Towards multidimensional genome annotation, Nat Rev Genet 2006;7:130-141.
3. Covert MW, Schilling CH, Famili I et al. Metabolic modeling of microbial strains in silico, Trends Biochem Sci 2001;26:179-186.
4. Henry CS, DeJongh M, Best AA et al. High-throughput generation, optimization and analysis of genome-scale metabolic models, Nat Biotechnol 2010;28:977-982.
5. Terzer M, Maynard ND, Covert MW et al. Genome-scale metabolic networks, Wiley Interdiscip Rev Syst Biol Med 2009;1:285-297.
6. Ruppin E, Papin JA, de Figueiredo LF et al. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks, Curr Opin Biotechnol 2010;21:502-510.
7. Feist AM, Herrgard MJ, Thiele I et al. Reconstruction of biochemical networks in microorganisms, Nat Rev Microbiol 2009;7:129-143.
8. Struhl K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes, Cell 1999;98:1-4.
9. Nudler E, Mironov AS. The riboswitch control of bacterial metabolism, Trends Biochem Sci 2004;29:11-17.
10. Mironov AS, Gusarov I, Rafikov R et al. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria, Cell 2002;111:747-756.
11. Wagner EG, Simons RW. Antisense RNA control in bacteria, phages, and plasmids, Annu Rev Microbiol 1994;48:713-742.
12. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs, Nat Rev Genet 2007;8:93-103.
13. Covert MW, Schilling CH, Palsson B. Regulation of gene expression in flux balance models of metabolism, J Theor Biol 2001;213:73-88.
14. Covert MW, Knight EM, Reed JL et al. Integrating high-throughput and computational data elucidates bacterial networks, Nature 2004;429:92-96.
15. Yoon H, McDermott JE, Porwollik S et al. Coordinated regulation of virulence during systemic infection of Salmonella enterica serovar Typhimurium, PLoS Pathog 2009;5:e1000306.

16. Herrgard MJ, Lee BS, Portnoy V et al. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae, Genome Res 2006;16:627-635.
17. Friedman N. Inferring cellular networks using probabilistic graphical models, Science 2004;303:799-805.
18. Gardner TS, di Bernardo D, Lorenz D et al. Inferring genetic networks and identifying compound mode of action via expression profiling, Science 2003;301:102-105.
19. Segal E, Shapira M, Regev A et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat Genet 2003;34:166-176.
20. Tegner J, Yeung MK, Hasty J et al. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling, Proc Natl Acad Sci U S A 2003;100:5944-5949.
21. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression, Proc Natl Acad Sci U S A 2002;99:6163-6168.
22. Mwangi MM, Siggia ED. Genome wide identification of regulatory motifs in Bacillus subtilis, BMC Bioinformatics 2003;4:18.
23. Rodionov DA, Dubchak I, Arkin A et al. Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria, Genome Biol 2004;5:R90.
24. Rodionov DA, Dubchak IL, Arkin AP et al. Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks, PLoS Comput Biol 2005;1:e55.
25. Babu MM, Lang B, Aravind L. Methods to reconstruct and compare transcriptional regulatory networks, Methods Mol Biol 2009;541:163-180.
26. Edwards JS, Covert M, Palsson B. Metabolic modelling of microbes: the flux-balance approach, Environ Microbiol 2002;4:133-140.
27. Shlomi T, Eisenberg Y, Sharan R et al. A genome-scale computational study of the interplay between transcriptional regulation and metabolism, Mol Syst Biol 2007;3:101.
28. Covert MW, Palsson BO. Constraints-based models: regulation of gene expression reduces the steady-state solution space, J Theor Biol 2003;221:309-325.
29. Yizhak K, Benyamini T, Liebermeister W et al. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model, Bioinformatics 2010;26:i255-260.
30. van Berlo RJP, de Ridder D, Daran JM et al. Predicting metabolic fluxes using gene expression differences as constraints, Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2011;8:206-216.
31. Young RA. Biomedical discovery with DNA arrays, Cell 2000;102:9-15.
32. Eisenstein M. Microarrays: quality control, Nature 2006;442:1067-1070.
33. Edgar R. Challenge of choosing right level of microarray detail, Nature 2006;443:394.
34. Brazma A, Hingamp P, Quackenbush J et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, Nature genetics 2001;29:365.
35. Microarray standards at last, Nature 2002;419:323.

36. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res 2002;30:207-210.
37. Brazma A, Parkinson H, Sarkans U et al. ArrayExpress--a public repository for microarray gene expression data at the EBI, Nucleic Acids Res 2003;31:68-71.
38. Velculescu VE, Zhang L, Vogelstein B et al. Serial analysis of gene expression, Science 1995;270:484-487.
39. Velculescu VE, Zhang L, Zhou W et al. Characterization of the yeast transcriptome, Cell 1997;88:243-251.
40. Ren B, Robert F, Wyrick JJ et al. Genome-wide location and function of DNA binding proteins, Science 2000;290:2306-2309.
41. Iyer VR, Horak CE, Scafe CS et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, Nature 2001;409:533-538.
42. Roh TY, Ngau WC, Cui K et al. High-resolution genome-wide mapping of histone modifications, Nat Biotechnol 2004;22:1013-1016.
43. Kim TH, Ren B. Genome-wide analysis of protein-DNA interactions, Annu Rev Genomics Hum Genet 2006;7:81-102.
44. Johnson DS, Mortazavi A, Myers RM et al. Genome-wide mapping of in vivo protein-DNA interactions, Science 2007;316:1497-1502.
45. Nagalakshmi U, Wang Z, Waern K et al. The transcriptional landscape of the yeast genome defined by RNA sequencing, Science 2008;320:1344-1349.
46. Mortazavi A, Williams BA, McCue K et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat Methods 2008;5:621-628.
47. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet 2009;10:57-63.
48. Herring CD, Raffaelle M, Allen TE et al. Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays, J Bacteriol 2005;187:6166-6174.
49. Wade JT, Roa DC, Grainger DC et al. Extensive functional overlap between sigma factors in Escherichia coli, Nat Struct Mol Biol 2006;13:806-814.
50. Grainger DC, Hurd D, Goldberg MD et al. Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome, Nucleic Acids Res 2006;34:4642-4652.
51. Perkins TT, Kingsley RA, Fookes MC et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi, PLoS Genet 2009;5:e1000569.
52. Croucher NJ, Thomson NR. Studying bacterial transcriptomes using RNA-seq, Current opinion in microbiology 2010.
53. Price MN, Deutschbauer AM, Kuehl JV et al. Evidence-based annotation of transcripts and proteins in the sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough, J Bacteriol 2011;193:5716-5727.
54. Barrett T, Troup DB, Wilhite SE et al. NCBI GEO: archive for high-throughput functional genomic data, Nucleic Acids Res 2009;37:D885-890.
55. Faith JJ, Driscoll ME, Fusaro VA et al. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata, Nucleic Acids Res 2008;36:D866-870.
56. Faith JJ, Hayete B, Thaden JT et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, PLoS Biol 2007;5:e8.

57. Overbeek R, Begley T, Butler RM et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, Nucleic Acids Res 2005;33:5691-5702.
58. Gama-Castro S, Salgado H, Peralta-Gil M et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units), Nucleic Acids Res 2011;39:D98-105.
59. Keseler IM, Collado-Vides J, Santos-Zavaleta A et al. EcoCyc: a comprehensive database of Escherichia coli biology, Nucleic Acids Res 2011;39:D583-590.
60. Sierro N, Makita Y, de Hoon M et al. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information, Nucleic Acids Res 2008;36:D93-96.
61. Okuda S, Yoshizawa AC. ODB: a database for operon organizations, 2011 update, Nucleic Acids Res 2011;39:D552-555.
62. Kazakov AE, Cipriano MJ, Novichkov PS et al. RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes, Nucleic Acids Res 2007;35:D407-412.
63. Novichkov PS, Laikova ON, Novichkova ES et al. RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes, Nucleic Acids Res 2010;38:D111-118.
64. De Smet R, Marchal K. Advantages and limitations of current network inference methods, Nat Rev Microbiol 2010;8:717-729.
65. Bansal M, Belcastro V, Ambesi-Impiombato A et al. How to infer gene networks from expression profiles, Mol Syst Biol 2007;3:78.
66. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks, Nat Rev Mol Cell Biol 2008;9:770-780.
67. Madan Babu M, Teichmann SA, Aravind L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks, J Mol Biol 2006;358:614-633.
68. Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in Escherichia coli, Nucleic Acids Res 2003;31:1234-1244.
69. Teichmann SA, Babu MM. Gene regulatory network growth by duplication, Nat Genet 2004;36:492-496.
70. Gelfand MS. Evolution of transcriptional regulatory networks in microbial genomes, Curr Opin Struct Biol 2006;16:420-429.
71. Lozada-Chavez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution, Nucleic Acids Res 2006;34:3434-3445.
72. Overbeek R, Fonstein M, D'Souza M et al. The use of gene clusters to infer functional coupling, Proc Natl Acad Sci U S A 1999;96:2896-2901.
73. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements, Nat Genet 2001;29:153-159.
74. Bar-Joseph Z, Gerber GK, Lee TI et al. Computational discovery of gene modules and regulatory networks, Nat Biotechnol 2003;21:1337-1342.
75. Alkema WB, Lenhard B, Wasserman WW. Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus, Genome Res 2004;14:1362-1373.
76. Rodionov DA. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria, Chem Rev 2007;107:3467-3497.

77. Tompa M, Li N, Bailey TL et al. Assessing computational tools for the discovery of transcription factor binding sites, Nature Biotechnology 2005;23:137-144.

78. Roth FP, Hughes JD, Estep PW et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation, Nature Biotechnology 1998;16:939-945.

79. Chen X, Guo LQ, Fan ZC et al. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data, Bioinformatics 2008;24:1121-1128.

80. Chen X, Guo L, Fan Z et al. Learning position weight matrices from sequence and expression data, Comput Syst Bioinformatics Conf 2007;6:249-260.

81. Tan K, Moreno-Hagelsieb G, Collado-Vides J et al. A comparative genomics approach to prediction of new members of regulons, Genome Research 2001;11:566-584.

82. Ravcheev DA, Rakhmaninova AB, Mironov AA et al. Comparative genomics analysis of nitrate and nitrite respiration in gamma proteobacteria, Molecular Biology 2005;39:832-846.

83. Doroshchuk NA, Gelfand MS, Rodionov DA. Regulation of nitrogen metabolism in gram-positive bacteria, Molecular Biology 2006;40:829-836.

84. Novichkov PS, Rodionov DA, Stavrovskaya ED et al. RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach, Nucleic Acids Res 2010;38:W299-307.

85. Rodionov DA, Novichkov PS, Stavrovskaya ED et al. Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the Shewanella genus, BMC Genomics 2011;12 Suppl 1:S3.

86. Suvorova IA, Tutukina MN, Ravcheev DA et al. Comparative genomic analysis of the hexuronate metabolism genes and their regulation in gammaproteobacteria, J Bacteriol 2011;193:3956-3963.

87. Leyn SA, Li X, Zheng Q et al. Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in Shewanella oneidensis, J Biol Chem 2011;286:35782-35794.

88. Oberto J. FITBAR: a web tool for the robust prediction of prokaryotic regulons, BMC Bioinformatics 2010;11:554.

89. McCue L, Thompson W, Carmack C et al. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes, Nucleic Acids Res 2001;29:774-782.

90. Su Z, Olman V, Mao F et al. Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis, Nucleic Acids Res 2005;33:5156-5171.

91. Brazhnik P, de la Fuente A, Mendes P. Gene networks: how to put the function in genomics, TRENDS in Biotechnology 2002;20:467-472.

92. De Jong H. Modeling and simulation of genetic regulatory systems: a literature review, Journal of computational biology 2002;9:67-103.

93. Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference, Ann N Y Acad Sci 2007;1115:1-22.

94. Marbach D, Prill RJ, Schaffter T et al. Revealing strengths and weaknesses of methods for gene network inference, Proc Natl Acad Sci U S A 2010;107:6286-6291.

95. Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges, Ann N Y Acad Sci 2009;1158:159-195.

96. Hache H, Lehrach H, Herwig R. Reverse engineering of gene regulatory networks: a comparative study, EURASIP J Bioinform Syst Biol 2009:617281.

97. Michoel T, De Smet R, Joshi A et al. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks, BMC Syst Biol 2009;3:49.

98. Elati M, Rouveirol C. Unsupervised Learning for Gene Regulation Network Inference from Expression Data: A Review, Algorithms in Computational Molecular Biology 2011:955-978.

99. Cloots L, Marchal K. Network-based functional modeling of genomics, transcriptomics and metabolism in bacteria, Curr Opin Microbiol 2011;14:599-607.

100. Cantone I, Marucci L, Iorio F et al. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches, Cell 2009;137:172-181.

101. Gustafsson M, Hornquist M, Lundstrom J et al. Reverse engineering of gene networks with LASSO and nonlinear basis functions, Ann N Y Acad Sci 2009;1158:265-275.

102. di Bernardo D, Thompson MJ, Gardner TS et al. Chemogenomic profiling on a genomewide scale using reverse-engineered gene networks, Nature Biotechnology 2005;23:377-383.

103. Butte AJ, Kohane IS (2000), 'Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements', pp. 418-429.

104. Meyer PE, Kontos K, Lafitte F et al. Information-theoretic inference of large transcriptional regulatory networks, EURASIP J Bioinform Syst Biol 2007:79879.

105. Yu J, Smith VA, Wang PP et al. Advances to Bayesian network inference for generating causal networks from observational biological data, Bioinformatics 2004;20:3594-3603.

106. Friedman N, Linial M, Nachman I et al. Using Bayesian networks to analyze expression data, Journal of computational biology 2000;7:601-620.

107. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures, Pac Symp Biocomput 1998:18-29.

108. Hache H, Wierling C, Lehrach H et al. (2007), 'Reconstruction and validation of gene regulatory networks with neural networks', pp. 319-324.

109. Grimaldi M, Jurman G, Visintainer R. Reverse Engineering Gene Networks with ANN: Variability in Network Inference Algorithms, Arxiv preprint arXiv:1009.4824 2010.

110. Rice JJ, Tu Y, Stolovitzky G. Reconstructing biological networks using conditional correlation analysis, Bioinformatics 2005;21:765-773.

111. Butte AJ, Kohane IS. Unsupervised knowledge discovery in medical databases using relevance networks, Journal of the American Medical Informatics Association 1999:711-715.

112. Ihmels J, Friedlander G, Bergmann S et al. Revealing modular organization in the yeast transcriptional network, Nature genetics 2002;31:370-377.
113. Bonneau R. Learning biological networks: from modules to dynamics, Nat Chem Biol 2008;4:658-664.
114. Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks, BMC Bioinformatics 2006;7.
115. Margolin AA, Nemenman I, Basso K et al. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, BMC Bioinformatics 2006;7.
116. Segal E, Pe'er D, Regev A et al. Learning module networks, Journal of Machine Learning Research 2005;6:557-588.
117. Joshi A, De Smet R, Marchal K et al. Module networks revisited: computational assessment and prioritization of model predictions, Bioinformatics 2009;25:490-496.
118. Mordelet F, Vert JP. SIRENE: supervised inference of regulatory networks, Bioinformatics 2008;24:i76-82.
119. Huynh-Thu VA, Irrthum A, Wehenkel L et al. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods, PLoS One 2010;5.
120. Bonneau R, Reiss DJ, Shannon P et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo, Genome Biol 2006;7:R36.
121. Lemmens K, De Bie T, Dhollander T et al. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli, Genome Biol 2009;10:R27.
122. Ernst J, Beg QK, Kay KA et al. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli, PLoS Comput Biol 2008;4:e1000044.
123. You ZH, Yin Z, Han K et al. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network, BMC Bioinformatics 2010;11:343.
124. Cerulo L, Elkan C, Ceccarelli M. Learning gene regulatory networks from only positive and unlabeled data, BMC Bioinformatics 2010;11:228.
125. Marbach D, Schaffter T, Mattiussi C et al. Generating realistic in silico gene networks for performance assessment of reverse engineering methods, J Comput Biol 2009;16:229-239.
126. Prill RJ, Marbach D, Saez-Rodriguez J et al. Towards a rigorous assessment of systems biology models: the DREAM3 challenges, PLoS One 2010;5:e9202.
127. Greenfield A, Madar A, Ostrer H et al. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models, PLoS One 2010;5:e13397.
128. Basso K, Margolin AA, Stolovitzky G et al. Reverse engineering of regulatory networks in human B cells, Nat Genet 2005;37:382-390.
129. Castro-Melchor M, Charaniya S, Karypis G et al. Genome-wide inference of regulatory networks in Streptomyces coelicolor, BMC Genomics 2010;11:578.
130. Glass K, Ott E, Losert W et al. Implications of functional similarity for gene regulatory interactions, J R Soc Interface 2012.

131. Yoon H, Ansong C, McDermott JE et al. Systems analysis of multiple regulator perturbations allows discovery of virulence factors in Salmonella, Bmc Systems Biology 2011;5.

132. Watkinson J, Liang KC, Wang XD et al. Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information, Challenges of Systems Biology: Community Efforts to Harness Biological Complexity 2009;1158:302-313.

133. Anastassiou D. Computational analysis of the synergy among multiple interacting genes, Molecular Systems Biology 2007;3.

134. Bonneau R, Reiss DJ, Shannon P et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo, Genome Biology 2006;7.

135. Madar A, Greenfield A, Vanden-Eijnden E et al. DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator, PLoS One 2010;5.

136. Yip KY, Alexander RP, Yan KK et al. Improved Reconstruction of In Silico Gene Regulatory Networks by Integrating Knockout and Perturbation Data, PLoS One 2010;5.

137. Geurts P, Irrthum A, Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology, Molecular Biosystems 2009;5:1593-1605.

138. Ernst J, Beg QK, Kay KA et al. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli, Plos Computational Biology 2008;4.

139. Zwir I, Huang H, Groisman EA. Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation, Bioinformatics 2005;21:4073-4083.

140. Lemmens K, De Bie T, Dhollander T et al. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli, Genome Biology 2009;10.

141. Balaji S, Babu MM, Aravind L. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E-coil, Journal of Molecular Biology 2007;372:1108-1122.

142. Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods, Bioinformatics 2011;27:2263-2270.

143. Narendra V, Lytkin NI, Aliferis CF et al. A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks, Genomics 2011;97:7-18.

144. Varner JD. Large-scale prediction of phenotype: Concept, Biotechnology and Bioengineering 2000;69:664-678.

145. Palsson B. The challenges of in silico biology, Nature Biotechnology 2000;18:1147-1150.

146. Varma A, Palsson BO. Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use, Bio-Technology 1994;12:994-998.

147. Covert MW, Schilling CH, Palsson B. Regulation of gene expression in flux balance models of metabolism, Journal of Theoretical Biology 2001;213:73-88.

148. Tomita M. Whole-cell simulation: a grand challenge of the 21st century, TRENDS in Biotechnology 2001;19:205-210.
149. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: Evaluating the consequences of constraints, Nature Reviews Microbiology 2004;2:886-897.
150. Papin JA, Hunter T, Palsson BO et al. Reconstruction of cellular signalling networks and analysis of their properties, Nature Reviews Molecular Cell Biology 2005;6:99-111.
151. Feist AM, Herrgard MJ, Thiele I et al. Reconstruction of biochemical networks in microorganisms, Nature Reviews Microbiology 2009;7:129-143.
152. Covert MW, Xiao N, Chen TJ et al. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli, Bioinformatics 2008;24:2044-2050.
153. Lee JM, Gianchandani EP, Eddy JA et al. Dynamic analysis of integrated signaling, metabolic, and regulatory networks, Plos Computational Biology 2008;4.
154. Ovacik MA, Androulakis IP. On the Potential for Integrating Gene Expression and Metabolic Flux Data, Current Bioinformatics 2008;3:142-148.
155. Tenazinha N, Vinga S. A Survey on Methods for Modeling and Analyzing Integrated Biological Networks, Ieee-Acm Transactions on Computational Biology and Bioinformatics 2011;8:943-958.
156. Machado D, Costa RS, Rocha M et al. Modeling formalisms in Systems Biology, AMB Express 2011;1:45.
157. Price ND, Papin JA, Schilling CH et al. Genome-scale microbial in silico models: the constraints-based approach, Trends Biotechnol 2003;21:162-169.
158. Llaneras F, Pico J. Stoichiometric modelling of cell metabolism, J Biosci Bioeng 2008;105:1-11.
159. Price ND, Famili I, Beard DA et al. Extreme pathways and Kirchhoff's second law, Biophys J 2002;83:2879-2882.
160. Edwards JS, Covert M, Palsson B. Metabolic modelling of microbes: the flux-balance approach, Environmental Microbiology 2002;4:133-140.
161. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis, Curr Opin Biotechnol 2003;14:491-496.
162. Shlomi T, Eisenberg Y, Sharan R et al. A genome-scale computational study of the interplay between transcriptional regulation and metabolism, Molecular Systems Biology 2007;3.
163. Covert MW, Palsson BO. Transcriptional regulation in constraints-based metabolic models of Escherichia coli, J Biol Chem 2002;277:28058-28064.
164. Hohmann S. Osmotic stress signaling and osmoadaptation in Yeasts, Microbiology and Molecular Biology Reviews 2002;66:300-+.
165. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis, Proceedings of the National Academy of Sciences of the United States of America 2010;107:17845-17850.
166. van Berlo RJP, de Ridder D, Daran JM et al. Predicting Metabolic Fluxes Using Gene Expression Differences As Constraints, Ieee-Acm Transactions on Computational Biology and Bioinformatics 2011;8:206-216.
167. Leek JT, Scharpf RB, Bravo HC et al. Tackling the widespread and critical impact of batch effects in high-throughput data, Nat Rev Genet 2010;11:733-739.

168. Otero JM, Panagiotou G, Olsson L. Fueling industrial biotechnology growth with bioethanol, Adv Biochem Eng Biotechnol 2007;108:1-40.
169. Bro C, Regenberg B, Forster J et al. In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production, Metab Eng 2006;8:102-111.
170. Vilaca P, Rocha I, Rocha M. A computational tool for the simulation and optimization of microbial strains accounting integrated metabolic/regulatory information, Biosystems 2011;103:435-441.
171. Kim J, Reed JL. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains, Bmc Systems Biology 2010;4.
172. Banta S, Vemula M, Yokoyama T et al. Contribution of gene expression to metabolic fluxes in hypermetabolic livers induced through burn injury and cecal ligation and puncture in rats, Biotechnology and Bioengineering 2007;97:118-137.
173. Schilling O, Frick O, Herzberg C et al. Transcriptional and metabolic responses of Bacillus subtilis to the availability of organic acids: Transcription regulation is important but not sufficient to account for metabolic adaptation, Applied and Environmental Microbiology 2007;73:499-507.
174. Shlomi T, Cabili MN, Herrgard MJ et al. Network-based prediction of human tissue-specific metabolism, Nature Biotechnology 2008;26:1003-1010.
175. Willenbrock H, Ussery DW. Chromatin architecture and gene expression in Escherichia coli, Genome Biology 2004;5.
176. Casadesus J, Low D. Epigenetic gene regulation in the bacterial world, Microbiology and Molecular Biology Reviews 2006;70:830-+.
177. Rodionov DA, Gelfand MS, Todd JD et al. Computational reconstruction of iron- and manganese-responsive transcriptional networks in alpha-proteobacteria, Plos Computational Biology 2006;2:1568-1585.
178. Moreno-Campuzano S, Janga SC, Perez-Rueda E. Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes - a genomic approach, BMC Genomics 2006;7.
179. Price MN, Dehal PS, Arkin AP. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli, Genome Biology 2008;9.
180. Balleza E, Lopez-Bojorquez LN, Martinez-Antonio A et al. Regulation by transcription factors in bacteria: beyond description, Fems Microbiology Reviews 2009;33:133-151.
181. Bornholdt S. Boolean network models of cellular regulation: prospects and limitations, Journal of the Royal Society Interface 2008;5:S85-S94.
182. Karr JR, Sanghvi JC, Macklin DN et al. A whole-cell computational model predicts phenotype from genotype, Cell 2012;150:389-401.
183. Barrett T, Troup DB, Wilhite SE et al. NCBI GEO: archive for functional genomics data sets--10 years on, Nucleic Acids Res 2011;39:D1005-1010.
184. Barrett T, Troup DB, Wilhite SE et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update, Nucleic Acids Res 2007;35:D760-765.
185. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation, Bioinformatics 2007;23:127-128.
186. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy, Nucleic Acids Res 2011;39:W475-478.

187. Sherlock G, Hernandez-Boussard T, Kasarskis A et al. The Stanford Microarray Database, Nucleic Acids Res 2001;29:152-155.

188. Engelen K, Fu Q, Meysman P et al. COLOMBOS: access port for cross-platform bacterial expression compendia, PLoS One 2011;6:e20938.

189. Robison K, McGuire AM, Church GM. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome, J Mol Biol 1998;284:241-254.

190. Jacques PE, Gervais AL, Cantin M et al. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis, Bioinformatics 2005;21:2563-2565.

191. Pauling J, Rottger R, Tauch A et al. CoryneRegNet 6.0--Updated database content, new analysis methods and novel features focusing on community demands, Nucleic Acids Res 2012;40:D610-614.

192. Wu J, Zhao F, Wang S et al. cTFbase: a database for comparative genomics of transcription factors in cyanobacteria, BMC Genomics 2007;8:104.

193. Perez AG, Angarica VE, Vasconcelos AT et al. Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes, Nucleic Acids Res 2007;35:D132-136.

194. Krawczyk J, Kohl TA, Goesmann A et al. From Corynebacterium glutamicum to Mycobacterium tuberculosis--towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet, Nucleic Acids Res 2009;37:e97.

195. Pareja E, Pareja-Tobes P, Manrique M et al. ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms, BMC Microbiol 2006;6:29.

196. Wilson D, Charoensawan V, Kummerfeld SK et al. DBD--taxonomically broad transcription factor predictions: new content and functionality, Nucleic Acids Res 2008;36:D88-92.

197. Grote A, Klein J, Retter I et al. PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes, Nucleic Acids Res 2009;37:D61-65.

198. Huang HY, Chang HY, Chou CH et al. sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes, Nucleic Acids Res 2009;37:D150-154.

Figure Legends

Figure 1 –Technologies, tools, and resources for transcriptional regulatory network modeling and reconstruction.

Figure 2- Survey of the GEO database. (a) Types of expression profiling studies on the database [183]. (b) Number of series of experiments available from next-generation sequencing technologies [183]. (c) Percentage of data from bacteria in the entire database: from a total of 28,150 series of experiments only 2,196 represent bacterial organisms. (d) Most-represented bacteria on GEO. The organisms presented have at least a minimum of 43 series of experiments. Data for (c) and (d) were obtained with GEO tools [184] in April 2012.

Figure 3 – Comparison of bacterial genomes with expression data in GEO versus genomes with complete DNA sequences in the PubSEED [57]. The 20 bacterial families that contain genomes with expression data in GEO are arranged in a topological tree. For each family, the most abundantly sampled species in the PubSEED was picked to represent that family, and the alignment of their 16S sequences was used to reconstruct the bacterial family tree. The color coding of the tree nodes denotes the phyla they belong to. Most phyla contain only one family, with the exception of Cyanobacteria (3 families), Bacteroidetes (4 families), and Firmicutes (3 families). The last two phyla are especially overrepresented in terms of both sequenced genomes and expression data. The numbers on the right of each tree node denote the number of genomes with GEO expression data (566 in total) and the number of genomes present in the PubSEED (3,493 in total). Archea organisms were removed from this study since we aim to survey only bacterial genomes. In the horizontal bar plot, we show the fraction of each bacterial family for which expression data is available (in dark red). The tree was designed with the Interactive Tree of Life Tool [185, 186].

Figure 4 – TRN reconstruction methodologies. (a) Template network based methods. b) TFBS data based via regulatory *cis* elements. (c) *De novo* reverse engineering.

Figure 5 – Network inference methods classification. (a) Network node Module Based vs Direct Inference. (b) Supervised vs unsupervised. Supervised methods require a training set of previous known interactions.

Figure 6 - Stoichiometric modeling. The metabolic network is used to construct the stoichiometric matrix using mass balances of the metabolites. The constraints-based approach is used to impose constraints to the stoichiometric model. S.v = 0 – pseudo steady-state assumption; v > 0 – reversibility constraint; v < vmax – capacity constraint.

Figure 7 – Pathway-based and constraints-based methods for the analysis and simulation of integrated metabolic and regulatory networks. FBA (flux balance analysis; rFBA (regulatory FBA); SR-FBA (steady-state regulatory FBA); idFBA (integrated dynamic FBA); iFBA (integrated FBA); PROM (Probabilistic Regulation of Metabolism); IOMA (Integrative Omics- Metabolic Analysis); tFBA (transcriptional controlled FBA).

Table 1 – Gene expression repositories with bacterial transcriptional data.

Database	Main Features
GEO [36]	NCBI's database for expression data. Supports multiple expression studies platforms for all organisms. Browsing tools available.
ArrayExpress [37]	EBI's database for expression data. Data submitted by users and imported from GEO. Advanced queries and ontology-driven searches.
M3D [55]	Data uniformly normalized from Affymetrix microarrays for *Escherichia coli, Saccharomyces cerevisiae* and *Shewanella oneidensis*.
SMD [187]	Partially public database with data from around 60 organisms. *Escherichia coli, Mycobacterium tuberculosis* and *Streptomyces coelicor* are among the most represented microbes. Data analysis framework embedded.
COLOMBOS [188]	Cross-platform expression compendia for E. *coli*, *B. subtilis*, and *S. enterica* subspecies serovar Typhimurium. Provides tools for expression analysis and extraction of relevant information.

Table 2 – Databases with notable bacterial transcriptional data.

Database	Organism(s)	Main Features
Organism specific		
DBTBS [60]	*B. subtillis*	Compendium of regulatory data with promoters, TFs, TFBS, motifs and regulated operons
RegulonDB [58]	*E. coli*	Compendium of regulatory data, promoters, TFs, TFBS, transcription units, operons and regulatory network interactions.
EcoCyc [59]	*E. coli*	Comprehensive database with gene products, transcriptional, post-transcriptional data and operon organization
DPInteract [189]	*E. coli*	DNA binding proteins and binding site data.
MTBRegList [190]	*M. tuberculosis*.	TFBS and regulatory motifs
Organism class/family		
CoryneRegNet [191]	Corynebacteria	TF and regulatory networks
cTFbase [192]	Cyanobacteria	Putative TFs
TractorDB [193]	Gamma-proteobacteria	TFBS predictions
MycoRegNet [194]	Mycobacteria	TF and regulatory networks
Non-organism specific		
ExtraTrain [195]	Bacteria and Archea	Transcriptional data and extragenic regions
DBD [196]		TF predictions
RegTransBase [62]		Regulatory interactions from literature and TFBS
PRODORIC [197]	Bacteria	TFs, TFBSs, regulon lists, promoters, expression profiles
sRNAMap [198]		Small noncoding RNAs and regulators
ODB [61]		Known and putative operons
RegPrecise [63]		Regulon database

Table 3 – Methods for reverse engineering of gene regulatory networks from expression data.

		Inference Approach		**Semi / Supervised**	
Algorithm	**Modeling Approach**	DI*	MB**	Yes	No
ARACNE [115]	Mutual Information (MI)	X			X
CLR [56]		X			X
SA-CRL [132]		X			X
tlCLR [135]	+ MI		X		X
Inferelator [134]	ODE Model		X		X
Yip *et al*. [136]	+ Noise Model	X			X
GENIE3 [119]	Regression tress	X		X	
SEREND [138]	Logistic regression	X		X	
GPS [139]	Fuzzy Clustering		X		X
DISTILLER [140]	Association rules (itemsets)		X		X

*DI – Direct Inference | **MB – Module-Based